2 мая 2023 г.

Задания для семинара № 6

Тема: Регрессионный анализ

1. Рассмотрим базу данных binary, доступную при запуске команды

   `read.csv("https://stats.idre.ucla.edu/stat/data/binary.csv")`.

   Таблица содержит информацию о результатах 400 студентов: поступление в аспирантуру (1/0 - первый столбец), результаты GRE (graduate record exam), GPA (grade point average) и престиж учебного заведения, в котором он/она получали бакалаврское образование (1 — наивысший рейтинг, 4 — наиболее низкий).

   (a) При помощи инструментов weights of evidence и information value разделите переменные GRE и GPA на группы (не менее 4 групп и не более 12 групп) и определите, при каком делении различие между группами наиболее выражено.

   (b) Для тех же данных, сравните при помощи анализа ROC-кривых следующие два метода предсказания поступления в аспирантуру:

   - для каждой группы вычисляется среднее значение переменной admit, которое является оценкой вероятности поступления в аспирантуру для данной группы. После этого для каждого из 400 студентов предсказывается вероятность поступления в университет, равная вероятности поступления в соответствующей группе;
   - метод логистической регрессии.

2. Рассмотрим базу данных `mtcars`, содержащую технические характеристики 32 марок автомобилей, произведённых в 1973-74 годы, см. `https://stat.ethz.ch/R-manual/R-devel/library/datasets/html/mtcars.html`.

   Целью данного упражнения является построение регрессионной модели, объясняющей зависимость потребления топлива (первая пе-

ременная `mpg`) от других характеристик автомобиля (переменные 3–7).

(i) Постройте линейную регрессию, показывающую зависимость между потреблением топлива и переменными 3–7. Проанализируйте взаимосвязь между переменными 3–7 и выберите 2 из них, наилучшим образом объясняющие потребление топлива на основе линейной модели. Визуализируйте полученную зависимость на трёхмерном графике.

(ii) Для тех же переменных постройте регрессионную модель методом Loess. Для выбора оптимальной зависимости используйте обобщённый метод кросс-валидации и критерий Акаике. Сравните среднеквадратичные ошибки полученных моделей и линейной модели.

(iii) Постройте оценку Надарая-Ватсона для описания зависимости между потреблением топлива и мощностью двигателя (переменная 4). Используйте различные виды ядер (гауссовское ядро, ядро Епанечникова), разные методы выбора параметра (обобщённый метод кросс-валидации, критерий Акаике), разные модификации модели (кусочно-постоянные и кусочно-линейные функции). Выберите модель с наименьшей среднеквадратической ошибкой.